KETAKEE NIMAVAT

ketakeenimavat91@gmail.com | (646) 341-7591 | San Francisco

linkedin.com/in/ketakee-nimavat

PROFESSIONAL EXPERIENCE

YouTube

GenAI Infra Engineer **San Francisco, November 2023 -current**

- Designed and led the implementation of the GenAI pipeline, tailored to address various upcoming YouTube AI use cases.
- Collaborated extensively across teams to architect a plug-and-play LLM pipeline enabling customer teams to configure their RAG components seamlessly.

Youtube Voice (Expanding Voice Search Across YouTube) **San Francisco, 2022- 2023**

- Designed and implemented efficient query fulfillment flow migration (speech recognition -> intent recognition and disambiguation -> query fulfillment) that resulted in 2% cost reduction, significant latency gains and better user experience.
- Spearheaded the optimization of microservice architecture, significantly enhancing latency (5-25% latency improvements and 1-5% watch time gains) for voice requests across all platforms (Youtube TV, YouTube Music, YouTube app).
- Integrated podcast use cases (intent recognition + fulfillment) into Youtube products.
- Engineered backend support for a personalized voice-driven UI that drove up user engagement(~1% voice watch time gains)

Oracle (Senior Member of Technical Staff for AI Apps, HCM) **San Francisco, *2020-2022***

- Built end-to-end micro-services to scale the recruiting recommendation product 3x in terms of no. of features.
- Initiated and delivered projects that use vector-based semantic search and vector databases, with the goals of improving the relevance of recommendations, and performance.
- Improved developer productivity by automating testing, implementing monitoring, and developing various internal tools for the team to simplify Kubernetes fleet management and product monitoring in production. The automations improved testing time 5x (from 5hrs -> 1 hr).
- Redesigned the existing data ingestion, storage, model training architecture to deliver real-time data ingestion and machine learning results for data at scale (100s of GB/customer).
- Filed for 2 patents to improve the relevance of recommendations using machine learning and leverage feedback in ML models.
- Translated business and functional requirements into deliverables while working with integration teams and worked with various VPs from Fortune 500 clients (in beta) to help onboard the company and understand their needs. Often experimented with the tech stack while adding new features iteratively, ultimately expanding those needs to GA features.

ADVISORY ROLES

- Provided strategic guidance and technical expertise to startups, helping them leverage LLMs to address customer challenges.
- Bootstrapped "Break It Down," an LLM based productivity app, to 10k YoY users.
- Advised App Genius AI, a no-code LLM startup, on product-market fit and LLM applicability for their use cases which led to them landing 5 customers
- Spearheaded the technical direction of Callo (a creative collaboration platform) by architecting the software stack, implementing recommendation algorithms, enhancing user experience, and strategizing for B2B and B2C initiatives.

INTERNSHIPS

Oracle , Software Development Intern **San Francisco, Summer 2019**

- Used deep learning to predict a user's app usage patterns and build a personalized feature layout on the homepage using a combination of prediction, ranking and filtering mechanisms using tensorflow and redux. Implemented a POC using federated machine learning for on device finetuning and secure recommendations.

WotNot.io , NLP Intern **India, Winter 2018**

- Created services such as entity extraction, intent detection, local /global context management, and information retreival for no-code FAQ chatbots.

EDUCATION

Columbia University (*Masters, Computer Science)* **NYC 2019**

- *Courses:* Machine Learning, Natural Language Processing, Applied Deep Learning, Practical Deep Learning System Performance.
- I*nteresting Projects:* Word sense disambiguation model, Hate speech classifier, Software Bug classifier using LSTMs and BERT.

L .D. College of Engineering*(Bachelors, Computer Engineering).* **India, 2018**

SKILLS

- Python, C/C++, Java, Solr, R, SQL, Go
- **Tools:** LangChain, Llama, huggingFace, Haystack, TensorFlow, Tensorflowjs spaCy, NLTK, Gensim, MongoDB, keras, Prometheus, Kibana, Grafana, Kubernetes, Docker